

LLM CONTINUUM

Entretiens d'auto- confrontation

HEAL – ENSC



Objectif



Dans le cadre du projet CONTINUUM, des tests ont été réalisés sur les interfaces suivis d'entretiens d'auto-confrontations pour recueillir les impressions des utilisateurs concernant ces interfaces. Les entretiens d'auto-confrontation ont été transcrits afin de mieux comprendre les retours des utilisateurs.

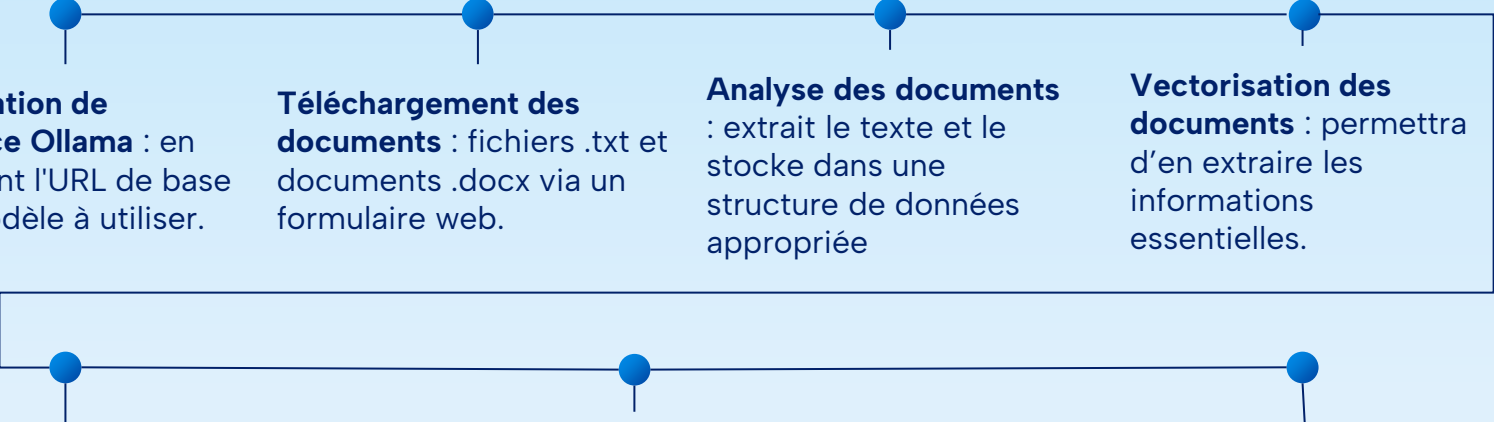
Pour analyser ces transcriptions de manière efficace, l'objectif de ce projet est de créer un code qui utilise l'outil Ollama pour tirer parti d'un modèle de Large Language Model (LLM). C'est à partir de ce modèle qu'a été réalisé un système de Retrieval Augmented Generation (RAG).

LLM : programme informatique intelligent qui peut comprendre et générer du langage humain de manière très sophistiquée

RAG (Retrieval Augmented Generation) : système informatique qui combine deux techniques : la recherche d'informations et la génération de texte. Il peut ainsi trouver des informations pertinentes à partir d'une grande quantité de données textuelles, puis utiliser ces informations pour générer du texte qui répond à une question ou qui développe un sujet donné.

Explication du code

ollama_langchain.py



Initialisation de l'instance Ollama : en spécifiant l'URL de base et le modèle à utiliser.

Téléchargement des documents : fichiers .txt et documents .docx via un formulaire web.

Analyse des documents : extrait le texte et le stocke dans une structure de données appropriée

Vectorisation des documents : permettra d'en extraire les informations essentielles.

Les utilisateurs **posent des questions** sur les documents téléchargés via un formulaire web.

Le modèle de RAG **trouve les réponses pertinentes** dans les documents vectorisés. Ollama fournit des fonctionnalités avancées de recherche et de génération de texte.

Affichage des réponses : affichées sur une page web dédiée, permettant aux utilisateurs de consulter les réponses à leurs questions.

Explication du code

index.html

Page d'accueil, renvoyant directement sur la page upload.html



upload.html

Page permettant de charger et vectoriser les documents .txt et .docs



responses.html

Pages affichant la réponse du RAG



ask_question.html

Page permettant de poser des questions sur les documents qui ont été chargés et vectorisés précédemment

Améliorations



- Actuellement, les réponses ne sont pas précises, elles restent vagues.
- Les réponses sont parfois en français, parfois en anglais.
- Il serait pertinent de pouvoir enregistrer les documents déjà vectoriser, pour ne pas avoir à le faire à nouveau.
- Il serait avantageux d'enregistrer automatiquement les différentes réponses pour pouvoir les consulter à nouveau.